Towards Enabling Critical mMTC: A Review of URLLC Within mMTC

SHIVA RAJ POKHREL[1], (Member, IEEE), **JIE DING**[1], **JIHONG PARK**[1], (Member, IEEE), **OK -SUN PARK**[2], (Member, IEEE), AND **JINHO CHOI**[1], (Senior Member, IEEE)

[1]School of IT, Deakin University, Geelong, VIC 3125, Australia
[2]Electronics and Telecommunications Research Institute, Daejeon 34129, South Korea

Corresponding author: Shiva Raj Pokhrel (shiva.pokhrel@deakin.edu.au)

ABSTRACT Massive machine-type communication (mMTC) and ultra-reliable and low-latency communication (URLLC) are two key service types in the fifth-generation (5G) communication systems, pursuing scalability and reliability with low-latency, respectively. These two extreme services are envisaged to agglomerate together into *critical mMTC* shortly with emerging use cases (e.g., wide-area disaster monitoring, wireless factory automation), creating new challenges to designing wireless systems beyond 5G. While conventional network slicing is effective in supporting a simple mixture of mMTC and URLLC, it is difficult to simultaneously guarantee the reliability, latency, and scalability requirements of critical mMTC (e.g., < 4ms latency, 10^6 devices/km^2 for factory automation) with limited radio resources. Furthermore, recently proposed solutions to scalable URLLC (e.g., machine learning aided URLLC for driverless vehicles) are ill-suited to critical mMTC whose machine type users have minimal energy budget and computing capability that should be (tightly) optimized for given tasks. To this end, our paper aims to characterize promising use cases of critical mMTC and search for their possible solutions. To this end, we first review the state-of-the-art (SOTA) technologies for separate mMTC and URLLC services and then identify key challenges from conflicting SOTA requirements, followed by potential approaches to prospective critical mMTC solutions at different layers.

INDEX TERMS Ultra reliable low latency communication (URLLC), massive machine type communications (mMTC), critical mMTC, 5G, beyond 5G.

I. INTRODUCTION

With the growing presence of enhanced wireless networks, ubiquitous infrastructure and vertical Internet of Things (IoT) domains (e.g., industry automation, autonomous vehicles), the evolution towards a comprehensive enabling platform for intelligent and tightly connected societies has already been evident. Such a platform could potentially consolidate the ever-increasing IoT applications, smart factory processes, public and businesses connectives jointly to uplift the overall quality of our daily lives. There are competing advances in both areas: i) licensed cellular systems, such as narrow-band IoT (NB-IoT) and long term evolution for machine-type communications (LTE-M), and ii) unlicensed bands such as LoRa and Zigbee, to support the connectivity of the vertical IoT

The associate editor coordinating the review of this manuscript and approving it for publication was Noor Zaman.

domains. We focus on cellular IoT in this paper as cellular systems have been celebrated due to their intrinsic benefits, including the provision of better quality of service and reliable coordination.

The ITU radiocommunication sector (ITU-R) has categorized 5G services into three broad classes, viz. massive machine type communications (mMTC), ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB). A simplified view of the identified goals for the third generation partnership project (3GPP) fifth generation (5G) can be outlined as follows:

○ yield up to 20 Gbps of eMBB speed and increase subscriber capacity by about hundred times of the current levels. We may not see this speed in practice, but we should see the downlink speeds uplifting from 100 Mbps to Gbps.

TABLE 1. Reliability and latency requirements of a few mMTC applications [1]–[3]. Observe that the delay requirements of 3D audio are quite stringent, which recreates sound in a much more natural way; it's closer to what we hear in real life.

Class	Reliability %	Delay ms	mMTC Apps. (see below)
1	$99.9 - 99.99999$	> 50	industry 4.0, vehicles, Haptics
2	$99.9 - 99.99999$	$10 - 50$	IIoT automation/orchestration
3	99.9 - 99.999	$2 - 10$	Vehicles, AR/VR, Drones
4	99.99999	2	Autonomous Cars, Haptics
5	99.999	1	Vehicles, Haptics
6	99.99999	2	Vehicles, Internet of Drones
7	99.999	2	AR/VR, Convoy of Drones
8	99.999	0.5	Cars, AR/VR Drones

mMTC Applications	Cases	Delay in ms	Reliability
Virtual Reality	Haptics	$0.5 - 2$	99.9 > 99.999
Augmented Reality	Video	$0.5 - 2$	> 99.999
Immersive Reality	3D-Audio	$0.5 - 2$	99.9
Industry 4.0	Haptics	highly-dynamic: 0.5 dynamic: 5 fixed: 50	99.9 > 99.999
	Video	2	> 99.999
	Audio	2	99.9
Autonomous Vehicles	Haptics	life critical: 0.5	99.9
	Sensor	dynamic: 5	> 99.999
	Video	dynamic: 5, fixed: 50	> 99.999
	Audio	2	99.9
Convoy/Swarm of Drones	Haptics	kinesthetic: 0.5 Tactile: 5	99.9 > 99.999
	Sensor	5	> 99.999
	Video	dynamic: 5, fixed: 50	> 99.999
	Audio	1-5	99.9
	GPS	5	99.9
Haptics communication (human touch feel)	Haptics	interaction: 0.5 Observation: 5	99.9 > 99.999
	Video	dynamic: 1, fixed: 5	> 99.999
	Audio	1-5	99.9

- deliver URLLC, i.e., $>$ 99.999% block error rate (BLER) and <1 milliseconds (ms) for some vertical IoT domains.
- enable massive connectivity, mMTC implementation and incorporation of low-power wide area network (LPWAN) requirements to strengthen IoT solutions.

These are all challenging goals, but have been mostly accomplished. Latency is still a concern in the optimal present levels of 10 milliseconds to 30 milliseconds range. The reliability and latency requirements for seamless operations of critical emerging mMTC applications (e.g., the advanced driving-assistance technologies for vehicles in a massive smart city applications, their autonomous cars and loss/delay sensitive industry 4.0 MTC for large factory automation) can barely be compromised. Therefore, our primary goal in this paper is to present a comprehensive review to summarize the current state of the art (SOTA) for *URLLC within mMTC*, refer to as *critical mMTC*, which is poorly understood in the literature.

To this end, one of the representative vertical domains is cellular-based smart manufacturing, which is characterized by the integration of IoT and related services in industrial automation. Standardized wireless network technologies and transmission protocols including 5G new radio (NR) are ineffective or unable to support low delay and ultra-high
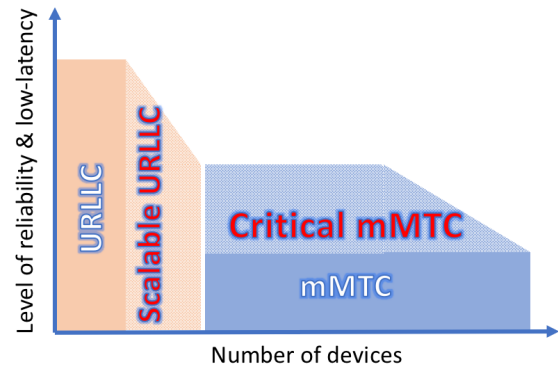

FIGURE 1. Future cellular IoT services.

reliability requirements of such emerging critical mMTC applications [4]. Intelligent integration of URLLC and mMTC has the potential to be a key enabler for the cellular-based industrial automation. As summarized in Table 1, in many such not-so-futuristic applications areas like i) augmented and virtual reality, ii) future industrial communication and control, iii) massive network of autonomous vehicles and sensors, iv) haptics, robotics, and tactile Internet in large factory networks; the utilization of extremely reliable and virtually zero delay wireless communications are important and indispensable along with the availability of massive connectivity [5], [6]. New standardization and design guidelines for *critical mMTC* services should be developed to fulfill their virtually error free and zero delay constraints.

In summary, both mMTC and URLLC need to set key parameters for higher layers to support heterogeneous emerging applications and services in the 5G systems, pursuing scalability and high reliability with low-latency, respectively. URLLC and mMTC are now envisaged to agglomerate together into *critical mMTC*, producing new challenges in designing systems beyond 5G. The factory automation application, for example, has evolved from wired massive sensors and actuators towards wireless intelligent machines and sensors for enabling flexible and smart manufacturing. A simple mixture of conventional mMTC and URLLC (e.g., radio access network slicing [7], [8]) cannot simultaneously guarantee their reliability, latency, and scalability over wireless links at the same level as perceived form wired factory environment (e.g., $<$ 4 ms latency,[1] 10^6 devices/km^2). Furthermore, generic solutions for scalable URLLC (e.g., machine learning (ML) aided URLLC for autonomous vehicles [9]) are ill-suited to critical mMTC [10], [11].

A. MOTIVATION OF CRITICAL mMTC
URLLC and mMTC are the two key service types supported by 5G, achieving reliability with low latency and scalability,

---

[1]Ericsson's measurement results showed that the network delay lower than 4 ms has no major impact on performance. It is because the robot's internal operation ends in a standard deviation of about 2 ms due to the internal sampling used in robot (https://www.ericsson.com/en/blog/2018/12/what-will-5g-bring-to-industrial-robotics).

TABLE 2. Survey and overview works in the areas of 5G URLLC and mMTC.

References	URLLC	mMTC	Scope
[4], [6], [11], [14]–[26]	[16]–[18], [21]–[26]	[4], [6], [15]–[20]	Enablers, challenges, applications
[27]–[31]	[27]–[30]	[30], [31]	Random Access (RA)
[32]–[34]	✓	✓	QoS, Data Analytics
[35]–[37]	✓	✓	Machine Learning
[38], [39]	—	✓	Highly Dense IoT

respectively. As shown in Fig. 1, it is envisaged that these two services will agglomerate into critical mMTC or scalable URLLC for beyond 5G with emerging use cases. In this subsection, we aim to justify the need of critical mMTC by highlighting key differences from existing similar notions.

- o **URLLC-mMTC Mixture** is the use case where URLLC and mMTC services are supported with separate physical resources by the same network. To support this heterogeneous system, the cellular resource configuration (i.e., numerology) has been re-designed [12], [13], and the mini-slot resources therein have been optimized using radio access network (RAN) slicing methods in an orthogonal or non-orthogonal resource allocation [8].

- o **Scalable URLLC** is targeted to support increased connections of MTC devices with a variety of URLLC requirements. This traffic type is *non-separable*, and thus cannot be supported using neither RAN slicing nor prioritizing methods used for the URLLC-mMTC mixture. Instead, ML based solutions have recently been proposed [9] in which prediction is adopted for reducing latency and improving reliability, without consuming radio resources.

- o **Critical mMTC** is targeted to support enhanced URLLC requirements for a fraction of massively connected MTC devices. It is possible to apply some techniques developed for the URLLC-mMTC mixture to critical mMTC. However, when a device can sometimes be a URLLC or non-URLLC device, resource configuration overhead may be overwhelming. Notwithstanding, it cannot be the same, especially for a large number of critical MTC devices that apply to scalable URLLC. Meanwhile, MTC devices generally lack powerful computing power and sufficient energy, so ML-based solutions for scalable URLLC are not suitable unless lightweight and energy-efficient ML architectures and algorithms are deployed.

In brief, critical mMTC is still a unique service type that has not been fully understood. This mandates our seeking for novel solutions to critical mMTC by unveiling its key characteristics and rethinking existing relevant approaches.

B. REVIEW OF RELATED SURVEY

Of particular relevance to this work is the survey and overview research works of URLLC and mMTC for 5G IoT. Table 2 provides a summary of references related to these areas categorized based on their themes. Several researchers

have provided a review of enabling approaches, ideas, challenges and applications of URLLC/mMTC in different IoT contexts [4], [6], [14]–[26], but the area of critical mMTC has been poorly reviewed in literature [16].

Furthermore, researchers have already summarized RA based schemes for the URLLC [27]–[30] and mMTC [30], [31] separately. A few other researches have covered a survey of the existing machine learning [35]–[37] and data analytics [32]–[34] approaches for URLLC and mMTC. A couple of surveys [38], [39] of highly densed IoT are relevant to the mMTC cellular IoT framework. Nevertheless, a review of enablers, challenges and techniques for the convergence of critical mMTC, our main theme in this paper, is a challenging and delicate task because of the complexity due to the heterogeneity in service and application constraints discussed earlier (recall Table 1).

C. CONTRIBUTIONS

While some of the recent survey/vision papers (see Tab. 2) considered various facets of mMTC and URLLC schemes, a thorough study of the research problems involved in supporting the large number of devices with a mix of mMTC and URLLC services in cellular IoT networks and a systematic evaluation of recent advances in 5G to resolve the challenges is missing in literature. As mentioned earlier, several challenges arise while incorporating URLLC and mMTC together into existing NBIoT-based cellular systems. Key concerns include the provisioning of QoS to heterogeneous mMTC systems, for example, fixing URLLC within mMTC, QoS-aware transmission scheduling, RAN congestion avoidance etc.

Overall, we highlight the main contributions of this paper as follows:

- o Major challenges faced by the current IoT (cellular) networks to support URLLC within a large number of mMTC devices are classified and possible future innovative ideas are noted along with essential elements, traffic characterization and potential vertical application domains.

- o Few ineffectiveness of the standard mMTC guidelines are acknowledged. With their application to critical mMTC systems, new ideas were proposed along with key features and channel control procedures for the evolving Cellular IoT standards.

- o Current higher-layer approaches for handling URLLC and mMTC separately in cellular IoT networks are reviewed along with insights on new concepts for overcoming critical mMTC specifications (jointly handling URLLC and mMTC) from higher-level perspectives.

- o Multiple research problems are identified and also some promising potential directions to promote future research efforts in the relevant fields are discussed.

D. STRUCTURE OF THE PAPER

To improve the readability of this paper, we provide a high level view of the paper structure in Fig. 2 and the
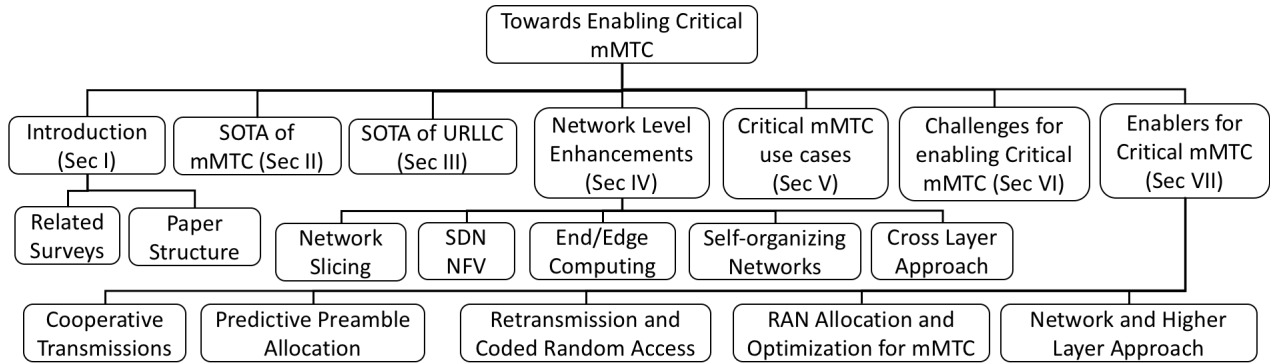
FIGURE 2. A high level view of topics and subtopics of the paper.

definitions of acronyms in Table 3. The remainder of the paper is structured as follows. Section II summarizes the SOTA of mMTC along with a discussion on NB-IoT, RAN and core network enhancements provisioning for massive connectivity. Section II also highlights some open issues for enabling massive connectivity. Section III reviews SOTA of URLLC with basic functionalities. Subsequently, Section IV highlights network level enhancements for both URLLC and mMTC. Section V presents important use cases of critical mMTC with some existing solutions. Section IV provides challenges towards enabling critical mMTC along with the difficulties at the network level. Finally, Section VII provides future directions towards potential critical mMTC enablers, and Section VIII concludes this paper.

II. STATE-OF-THE-ART OF mMTC

Most of mMTC use cases are characterized by a scenario where a large number of machine-type devices deployed in a wide area sporadically communicate without specific latency requirements, ignoring reliability guarantees. Supporting these use cases is commonly recast by meeting the following key performance indicators (KPIs) [40]:

- ○ Massive connection density of 10^6 devices per square kilometer in an urban environment;
- ○ Maximum coupling loss (MCL) up to 164 dB for wide coverage; and
- ○ Device battery lifetime over 10 years with a stored energy capacity of 5 Wh.

To achieve these mMTC KPIs, 3GPP has developed two radio technology standards built on the existing long term evolution (LTE) systems: LTE-M for lightweight mobile applications and voice-over-LTE (VoLTE); and NB-IoT for low-rate and wide-coverage applications. At ITU-R WP5D#32 meeting in July 2019, NB-IoT was officially recognized as a 5G candidate solution to meet the technical requirements of large-scale mMTC scenarios. Hereafter, we therefore focus on NB-IoT as a representative mMTC technology, and overview its key features as well as major enhancements.

A. FEATURES OF NB-IoT

NB-IoT is a radio technology proposed in 3GPP Release 13 towards supporting narrow-band LPWAN, a commercially successful wide-area mMTC application. Since NB-IoT is built upon LTE networks, NB-IoT can exploit LTE network hardware and resources, thereby reducing the deployment as well as operational costs [18]. At the physical layer, each NB-IoT carrier requires a minimum bandwidth of 180 kHz, which is equivalent to one LTE Physical Resource Block (PRB), and has three operational modes utilizing different amount of spectrum [41]. Specifically, NB-IoT can reserve: one or more LTE PRBs within an LTE carrier under the *in-band mode*; one or multiple global systems for mobile communications (GSM) carriers under the *stand-alone mode*; and only the spectrum within the guard-band of an LTE carrier under the *guard-band* mode.

1) MASSIVE CONNECTIVITY

To achieve massive connectivity, resource unit (RU) configurations based on subcarriers or tones are introduced for multiple access [42]. Specifically, single-carrier frequency-division multiple access (SC-FDMA) is applied to uplink, using either 3.75 kHz or 15 kHz subcarrier spacing. With a 3.75 kHz subcarrier spacing, NB-IoT can only allocate single-tone RU with a duration of 32 ms to different users. This mode is designed for power-constrained devices in a wide area. In addition, the reduced subcarrier bandwidth allows the simultaneous allocation of up to 48 devices in within one PRB. On the other hand, with a 15 kHz subcarrier spacing, either single-tone RU (8 ms) or multi-tone RUs (3, 6, or 12 tones) with a duration of 4 ms, 2 ms, and 1 ms, respectively, can be configured to different devices. For single-tone configurations, phase rotated $\pi/2$-binary phase shift keying (BPSK) or $\pi/4$-quadrature phase shift keying (QPSK) modulations can be used. For multi-tone configurations, only QPSK modulation is used [43]. For narrowband physical random access channel (NPRACH) of NB-IoT, a single-tone based preamble of 4 symbol groups with frequency hopping is designed [44]. In addition, as in LTE, orthogonal

TABLE 3. Definitions of abbreviations.

Abbreviations	Definitions
3GPP	Third Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
WUE	Wake Up Signal
BLER	Block Error Rate
BPSK	Binary Phase Shift Keying
BS	Base Station
CS	Compressive Sensing
DL	Downlink
DPSK	Differential Phase Shift Keying
DQPSK	Differential Quadrature Phase Shift Keying
FDMA	Frequency-Division Multiple Access
GSM	Global System for Mobile Communication
ISM	Industrial, Scientific, and Medical
ITU	International Telecommunication Union
IoT	Internet of Things
LPWAN	Low Power Wide Area Networks
LTE	Long-Term Evolution
LTE-A	Long-Term Evolution-Advance
LTE-M	Long-Term Evolution Machine Type Communications
ML	Machine Learning
MTC	Machine Type Communications
NB-IoT	Narrow-Band IoT
NFV	Network Function Virtualization
NOMA	Non-Orthogonal Multiple Access
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiplexing Access
OOK	On-Off Keying
OQPSK	Offset Quadrature Phase-Shift Keying
PDM	Power-Domain Multiplexing
PUSCH	Physical Uplink Shared Channel
NPRACH	Narrowband Physical Random Access Channel
PRB	Physical Resource Block
PSM	Power Saving Mode
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase-Shift Keying
QoS	Quality of Services
RA	Random Access
RAN	Random Access Network
RF	Radio Frequency
RL	Reinforcement Learning
RRC	Radio Resource Control
SC-FDMA	Single-carrier Frequency-Division Multiple Access
SDN	Software Defined Networking
SIC	Successive Interference Cancellation
SINR	Signal-to-Noise-and-Interference Ratio
SLA	Service Level Agreement
TDD	Time Division Duplex
TDMA	Time-Division Multiple Access
TTL	Time-To-Live
TTI	Transmission Time Interval
UAV	Unmanned Aerial Vehicles
UL	Uplink
V2X	Vehicle-to-Everything
WLAN	Wireless Local Area Networks
WPAN	Wireless Personal Area Networks
WAC	Wide Area Coverage
WSN	Wireless Sensor Networks
mMIMO	Massive Multiple-Input Multiple Output
eDRX	Expanded Discontinuous Reception

frequency-division multiple access (OFDMA) is used in the NB-IoT downlink with a 15 kHz subcarrier spacing only.

2) WIDE COVERAGE
In order to extend the coverage range in an open environment and compensate the penetration losses in a challenging indoor

space with highly reliable communication, NB-IoT targets up to 164 dB MCL, which is 20 dB coverage enhancement compared to GSM and general packet radio service (GPRS). To achieve this goal, except operating in narrow bandwidth, NB-IoT supports the approach of transmission repetitions (i.e., exploiting time diversity), by which the received signal-to-noise ratio (SNR) can be enhanced so that data could be decoded even when the signal power is much lower than the noise power [45]. In NB-IoT, the signal transmission can be repeated up to 128 and 2048 times in the uplink and downlink, respectively, and the number of repetitions depends on the coverage enhancement level required by devices along with the number of tones and subcarrier spacing [46].

3) LOW POWER CONSUMPTION
To lengthen battery life of devices, wake up signal (WUS) is used in NB-IoT to allow devices to avoid regularly paging checking and only start the procedure once WUS is received [47]. In addition, NB-IoT inherits two power saving techniques of LTE, i.e., power saving mode (PSM) and expanded discontinuous reception (eDRX) [48]. In particular, PSM allows a device registered on a network to turn off the functionalities of paging listening and link quality measurements to save energy. With eDRX, a device can negotiate with a network the time that it can fully turn off the receiving functionality for energy saving.

B. ENHANCEMENTS FOR NB-IoT
The continued evolution of NB-IoT is an important 3GPP activity. For instance, 3GPP Release 14 had enhancements in the form of support for higher data rates, multicast, positioning, a lower power class, and system access on non-anchor carriers [18]. 3GPP Release 15 included further enhancements in the form of support for improved latency, power consumption, measurement accuracy, cell range and load control. To extend the range of deployment options, it also specified small-cell and time division duplex (TDD) support for NB-IoT [47]. In 3GPP Release 16, additional enhancements and extensions for NB-IoT are specified to further improve the efficiency of network operation. Particularly, improvements on spectral efficiency for NB-IoT transmission and energy efficiency for NB-IoT devices are studied and proposed [12], which include: 1) enhanced mobile-terminated early-data transmission; 2) support for device-group WUS; and 3) improved uplink transmission using pre-configured resources in idle mode. Furthermore, 3GPP Release 16 also includes a common 2-step PRACH to decrease the latency and reduce additional control-signaling overhead. This is achieved by combining the preamble and the scheduled physical uplink shared channel (PUSCH) transmission into a single message from devices, known as MsgA. Then by combining the random-access respond and the contention resolution message into a single message (MsgB) from the BS to devices [12].

In addition to the enhancements by 3GPP, it has been conducted to analyze and improve the performance of NB-IoT

networks in a range of areas including: 1) NPRACH evaluation and enhancements [49]–[53]; 2) coverage analysis and enhancements [46], [54]–[56]; 3) energy efficiency enhancements [57]–[59]; and 4) co-channel interference analysis and mitigation [60]–[63].

C. RAN & NETWORK ENHANCEMENTS FOR MASSIVE CONNECTIVITY

3GPP has continued to expand the 5G infrastructure with new mMTC technologies for cellular IoT support, 5G LAN services, industry 4.0 time-aided networking and optimized access and relay backhaul [2]. It has also developed features for improving service-oriented architecture, enhancing flexible session management deployments, user plane and control functions, and supporting commercial services using location-based service architecture, enhancing self-organizing networks, smart dual connectivity, and aggregating dynamic carrier [2, Sec 4.2].[2]

1) 5G ENHANCED SUPPORT FOR VERTICAL IoT DOMAINS

Some of the key features identified by 3GPP relevant to our context are: enhanced support of LAN services (Vertical_LAN), 5G support for Cellular IoT evolution (5G_CIoT), support for Industrial IoT (NR_IIoT), architectural enhancements for advanced vehicle to everything (V2X) services (eV2XARC) and enablers for network automation and orchestrations [17], [26].

2) SERVICE BASED ARCHITECTURE ENHANCEMENTS

Service based enhancements scale and help for flexible implementation of proxy-based infrastructure and communication network through repository function and service [64]. The network functions and services defined can be implemented and used selectively by maintaining a set of services and a set of functionalities.

3) ENHANCED NETWORK AND RAN SLICING

The network slicing provides a mechanism for reallocation of access and mobility management along with session management and control functions [17]. It also facilitates independent and slice specific authentication and authorization per network slice. Starting from the granularity of network slicing per user or one slice for each services, we can identify how these slices are mapped and implemented at the RAN level in an efficient way in terms of radio resource consumption. We can start from the lower layers up to network function selection, configuration and chaining for each slice [65].

4) ENHANCED AUTOMATION AND ORCHESTRATION

There are advancements in the data collection and network analytic features specific to data collection based on the source and type of information to network analytics of mMTC, their automation and orchestration functions such as slice specific load balancing, patching, evaluation of

network performance indicators, user mobility and sustainable QoS [66]. With such enhancements, the network is capable of utilizing multiple data analytics, more importantly, analytics support are available at several levels.

5) ARCHITECTURE SUPPORT FOR TIME-BASED NETWORKING

New 5G architectural enhancements enable services to provide time synchronization of packet delivery in each hop of mMTC, to support time sensitive networking such as industrial automation [67]. 5G system can be easily integrated with an external network for desired level of services [68]. For example, a centralized model and with a specific subset of specifications and features can be enabled for such integration. Typical areas include periodic QoS guarantees, store and forward buffer, and logical bridging management for desired QoS mapping. Such a time-based approach includes a unique and even multiple working domains via a single RAN architecture [2].

III. STATE-OF-THE-ART OF 5G URLLC

URLLC aims to support mission-critical applications ranging from factory automation to vehicle safety control. In sharp contrast to conventional communication services pursuing the best-effort performance, these mission-critical applications commonly strictly require high reliability and low latency guarantees (e.g., $>$ 99.999% BLER with $<$ 1 ms user-plane latency [12]). In 3GPP specifications, new radio (NR) standards have been developed to enable URLLC. In this section, we provide an overview of key URLLC features and major enhancements in NR.

A. BASIC URLLC FUNCTIONALITY

Most of the basic URLLC functionalities in NR were filed within 3GPP Release 15 and finalized in September 2018. Thus, we focus on 3GPP Release 15, and elaborate the key functionalities provisioning URLLC service level agreement (SLA) in terms of latency and reliability [12].

1) FUNCTIONALITY FOR LOW LATENCY

Two main functionalities for low latency are explained as follows.

a: FLEXIBLE NUMEROLOGY AND FRAME STRUCTURE

NR supports scalable numerology and flexible framework to address heterogeneous configurations, deployment, and services [6]. In particular, subcarrier spacing of $2^n \times 15$ KHz ($n = 0, 1, 2$) can be supported for data channels at sub-6 GHz in NR. To achieve the low latency target of URLLC, it is necessary to use short transmission time interval (TTI). In light of this, large subcarrier spacing should be employed for URLLC transmission. For example, the TTI of 14 OFDM symbols with 15 kHz subcarrier spacing is 1 ms, while that with 60 kHz subcarrier spacing is reduced to 0.25 ms. Furthermore, NR also introduced the concept of mini-slot (consisting of only 2, 4, or 7 OFDM symbols) to further shorten TTI.

---

[2]https://www.3gpp.org/news-events/2122-tsn_v_lan.

For example, the TTI of a 2-symbol mini-slot with 60 kHz subcarrier spacing is only 35.7 μs, which facilitates to meet the low-latency goal.

In addition, other flexible framework designs such as frequent transmission opportunities, self-contained slot structure, shortening hybrid automatic repeat request (HARQ) roundtrip time, and traffic preemption for URLLC are also the important features to enable low-latency transmissions [1].

b: CONFIGURED-GRANT (GRANT-FREE) SCHEDULING

The conventional grant-based PRACH requires each device to transmit a scheduling request first and then wait for an uplink grant from the BS. The complex handshaking procedure between devices and BS can result in excessive signaling overhead and makes it difficult to meet the URLLC latency requirement of 1 ms or less in some scenarios. To reduce the access delay, configured-grant (grant-free) scheduling in 3GPP was proposed to allow devices to transmit uplink data without uplink grant on the configured resources [69]. The periodicity of configured resources can be as short as 2 symbols. The devices in grant-free scheduling can be dedicated devices with periodic traffic for non-contention based PRACH or a group of devices with aperiodic or sporadic traffic for contention based PRACH.

2) FUNCTIONALITY FOR HIGH RELIABILITY

Three important functionalities for high reliability are discussed in the following.

a: LOW SPECTRUM-EFFICIENT MCS/CQI

To guarantee reliable data transmission, optimal modulation and coding scheme (MCS) based on the channel quality indication (CQI) should be employed according to a look-up table [70]. In URLLC, reliability can be improved at the expense of low spectrum efficiency. To this end, one can either enlarge resource bandwidth or shorten TTI, which is of help to use low coding rate for reliability enhancement. Therefore, low spectral efficiency entries, e.g., QPSK with 1/8 coding rate, are considered in MCS table for URLLC.

b: MULTI-SLOT REPETITION

Repetition is a common approach to improve reliability as well as coverage. In URLLC, multi-slot repetition (2, 4, and 8 repetitions) can be used in data channels without waiting for any grant or retransmission feedback, which is also useful in URLLC for devices without sufficient time to provide or wait for HARQ feedback. Note that repetitions can also be applied to the control channels.

c: DIVERSITY EXPLOITATION

In URLLC, time diversity is not a viable solution for reliability enhancement as the packet cannot span over a long time due to a tight latency budget [19]. However, exploiting diversity in the frequency and spatial domains could be the ways to improve the reliability of URLLC. For example, frequency diversity could be achieved by frequency hopping and spatial diversity can be achieved by dual-connectivity or multi-connectivity of non-collocated BSs, or by multiple-input multiple-output (MIMO or massive MIMO) technologies.

B. ENHANCEMENTS FOR URLLC

3GPP Release 15 was revolutionary in terms of introducing a brand-new NR for URLLC, whereas Release 16 and 17 were evolutionary, targeting new verticals with tighter requirements by enhancing the capacity and operation of existing features [12]. In particular, Release 16 has expanded URLLC to the new verticals such as factory automation, transport industry, and electrical power distribution. To support more strict URLLC requirements, it further focused on the reliability enhancements on the control message transmission and latency reduction on HARQ feedback. Moreover, it studied the potential benefits of uplink inter device prioritization and multiplexing including uplink preemption and enhanced power control. It also supported multiple active configurations for configured-grant enhancements, so that different service traffic can be accommodated and quicker alignment for URLLC uplink transmissions can be made. In addition, the 2-step PRACH proposed in Release 16 can also be applied to URLLC. Release 17 will bring more new use cases to enable everything connected for 5G evolution systems in 2020 and on wards. Particularly, studies of URLLC in Release 17 will mainly focus on new emerging verticals and end-to-end performance of different applications, building enhancements to Release 16 features [69].

In addition to the studies and enhancements by 3GPP, several research works have also contributed to the analysis and enhancements for URLLC. For instance, in [71], a tractable approach is proposed to derive and analyze the latent access failure probability of URLLC device under three different grant-free RA schemes, namely reactive, K-repetition, and proactive, and showed that the proactive scheme can provide the lowest latent access failure probability under shorter latency constraints. In [69], the shortcomings of existing grant-free RA schemes are discussed for enabling URLLC, and two advanced grant-free RA schemes that go beyond 5G NR are proposed by taking advantage of non-orthogonal multiple access (NOMA). Furthermore, in [72], repetition for NOMA has been studied to lower outage probability.

IV. NETWORK LEVEL URLLC AND mMTC ENHANCEMENTS

While much attention has been focused on the physical and link layers, it is increasingly being realized that a wider redesign at network level is also essential to meet the specified requirements. It is also known that the existing cellular network architecture cannot support diversified services as network design fundamentals remains unchanged – largely based on conventional mobile broadband services. Thus, novel mobile computing frameworks and flexible network architectures are to be developed. These include approaches such as network slicing, software-defined networking (SDN),

network function virtualization (NFV), orchestration and self-organizing networks (SONs). In this section, we cover each of these approaches and their advancements for URLLC services.

A. DYNAMIC AND CONSTRAINED NETWORK SLICING

Network slicing has been a fundamental technique to harness diversified services simultaneously over the same physical infrastructure for future networks by enabling the deployment of multiple virtual domains atop a shared infrastructure [7], [8]. In scenarios like eMBB, mMTC, and URLLC, network slicing not only i) allows building multiple logical subnetworks with reserved resources enhancing the quality of service but also ii) avoids interruptions caused by other services. For example, to accommodate a number of critical mMTC (different) scenarios, or URLLC within mMTC use cases, we should be able to manipulate the network on the fly. For this vision to be realistic, network operators should be able to orchestrate their capabilities across several points in the infrastructure dynamically – *dynamic network slicing* [73], [74]. Furthermore, to satisfy critical mMTC requirements, we need a different approach to resource allocation among network slices for guaranteed and context aware slicing of resources– *constrained network slicing* [75].

The primary benefit to exploit such a dynamic network slicing for critical mMTC is that the network operators will then be able to plug and play with the slices timely to satisfy both URLLC and mMTC temporal services. For example, constrained dynamic network slicing can be used to manage the time varying network traffic, when an event occurs or disaster happens, the area get overwhelms with downlinks first as most devices download information about the event before or when it begins. However, after sometime, it will be opposite in the sense that the whole traffic shifts to the uplink because devices then start streaming live and uploading data/information about the event to the server for data analytics to generate next actions. To attain such a constraint and dynamic network slicing, the slices should be optimizing their resource allocations under loss and delay constraints and satisfy the URLLC requirements [76]. In particular, constrained dynamic optimization may help to guarantee that the slices are managed in a desired fashion and the network resources are shared to meet their SLAs depending on the applications' need at the higher layers.

More importantly, such slicing would be able to tailor the infrastructure by observing the time-evolving traffic pattern and network dynamics. As mentioned, network slicing has the potential to tackle and capitalize such scenario by utilizing the capabilities of NFV, orchestration, SDN and analytics [77]. However, strictly isolating and/or dynamic sharing of resources among multiple slices has been facing significant challenges because of the corresponding traffic fluctuations due to evolving users and varying channel conditions, therefore, it requires further investigations [8], [78]. Other intelligent solutions utilizing data-driven machine learning and artificial intelligence become crucial for several vertical

domains including but not limited to, more efficient manufacturing, urban computing and autonomous traffic settings [79].

B. SOFTWARE DEFINED NETWORKING AND NETWORK FUNCTION VIRTUALIZATION

SDN and NFV are the two successful network architecture techniques that support dynamic network slicing [68], [80]. Caballero *et al.* [80] illustrated via game-theoretic tools that simplification of scheduling and resource allocation in software-defined networking can be accomplished by separating the network control plane from the packet forwarding plane which offers dynamic network flow management. Additionally, network function visualization provides a high degree of programmability and flexibility by decoupling network functions from dedicated hardware devices. We have observed that integrating these two techniques can significantly provide efficient, scalable, and flexible network slicing service configurations for improving the overall performance for the network layer. In fact, it ameliorates both latency and reliability substantially [17]. It has been reported that software-defined networks can achieve an up to 75% performance improvement in end-to-end latency [5]. Ksentini *et al.* [81] proposed two-level medium-access-control scheduling framework for slicing by using dynamic slice management and established the QoS requirement for URLLC and mMTC. Zheng *et al.* [82] proposed constraint enabled resource slicing.

C. EDGE AND END NODE COMPUTING

Mobile edge/end node computing is another best solution that can reduce latency while processing tasks in URLLC with massive connectivity [76]. Mobile edge computing can be integrated with software-defined networks and network function virtualization to deal with such service disruption [83]. Improvement in resiliency and reduction in latency can be significantly observed when distributed, and virtualized networks are provisioned efficiently. Computation and transmit power can be significantly saved by optimizing the offloading process and required resource allocation in such mobile edge computing frameworks. However, the tradeoff between delay and computing power requires further investigations along the lines of [84]. Sood *et al.* [85] proposed end node computing and reported several challenges. Further extensive investigations are required to determine the best optimization techniques at the edge and end devices from the task offloading and scheduling prospective to provide a high quality of service in the URLLC of massive connectivity.

D. SELF-ORGANIZING NETWORKS

All of the aforementioned techniques, like network slicing, software-defined networks, and network function visualization improve network scalability and flexibility, however, the quality of service and perceived user experience may be compromised due to the integration of the techniques and complication in network management and

their parameter configuration. Therefore, it is critical to adopt self-organizing network management strategies, which could provide the desired optimization, distributed management, intelligence, and automation in such convergent network layer approaches [86]. In [87], a joint scheduling framework is proposed with traffic steering for eMBB and URLLC. However, to attain these, an autonomous intelligent service-aware SON is required for the such network with control functions at a network layer central unit, and various edge and core servers [88]. A new catalog-driven network management system that enables the smart deployment of service has already been proposed by the European Union's 5G Public-Private Partnership Project. Such self organization can be achieved by several machine learning techniques as outlined in [89].

E. CROSS-LAYER ENHANCEMENTS

Cross-layer design has the potential to improve both the end-to-end delay and the overall reliability significantly. This is feasible as each layer of the TCP/IP protocol stack is inherent and interdependent with the other layers. As an example, the transmission of packets, queuing in buffers, and delays in routing mechanisms all depend on the dynamics of the physical, link, and network layers, respectively. Better resource utilization can definitely be achieved by optimizing the delay components based on the end-to-end delay constraints. To this end, the investigation for *bandwidth or energy* optimization by adopting a cross-layer design is on real momentum [90], [91].

Collins and Cruz [92] introduced, first of all, the concept of cross-layer scheduling on the basis of the queueing and channel conditions in order to minimize average power usage under the average delay limit. Wang *et al.* [93] developed a cross-layer approach to manage data networking latency in wireless sensor networks, resulting in better energy efficiency. She *et al.* [94] adopted the cross-layer technique in-radio communication network for URLLCs (given delay is shorter than the width of the signal). More importantly, authors in [90], [91] studied cross-design approach for offloading and NOMA systems.

V. CRITICAL mMTC USE CASES AND EXISTING SOLUTIONS

Based on the overviews of mMTC and URLLC in the preceding sections, we summarize the SLAs of mMTC and URLLC in Table 4.

Given the distinct SLAs of mMTC and URLLC, in the this section we focus on possible critical mMTC uses cases wherein a massive number of non-critical MTC and critical MTC devices coexist, and discuss existing solutions to such use cases.

A. CRITICAL mMTC USE CASES

Critical MTC use cases, requirements, and traffic characteristics have been discussed in [1]–[3], as summarized in Table 5. These critical MTC devices are expected to coexist with

TABLE 4. Summary of requirement and specification disparity between mMTC and URLLC.

Specifications	mMTC	URLLC
Connection Density	Up to $10^6/\text{km}^2$	Comparably low
Power Consumption	Extreme low	Insensitive
End to End Latency	Insensitive	$1-10$ ms
Reliability	Typical BLER 10^{-1}	Up to BLER 10^{-9}
Payload Size	Small	Small to large
Bandwidth	Narrowband	Wideband
Numerology	3.75 KHz 15KHz	$2^n \times 15\text{KHz}$

non-critical MTC devices, forming the heterogeneous traffic of critical mMTC. 3GPP has worked on mMTC/URLLC as a new NR approach to meet the stringent reliability and delay needs of critical mMTC use cases, such as manufacturing and robotics. Such offerings could be complemented by adopting the new NR tailored to assist industrial wireless sensors. The NR based on numerology and SSB bandwidth, but complemented with enhancements for reduced complexity and low power consumption.

One exemplary use case of critical mMTC is studied in [95] from smart city perspective. In the smart city, a massive number of devices perform various tasks ranging from environmental and critical infrastructure monitoring to smart grids and industrial automation. When a triggering event occurs, a fraction of these devices become sensing devices that observe critical triggering information. These devices are grouped based on their functionalities and/or geographic locations, forming a critical MTC group while the remainder becomes non-grouped mMTC devices. A critical MTC group header is then assigned for measuring the urgency level of each event and communicating with a BS through URLLC.

Fitzgerald *et al.* [96] studied another critical mMTC use case is investigated in the context of industrial automation and the use of massive MIMO. In this scenario, the traffic is categorized into control and alarm traffic classes. Control traffic encompasses the transmissions between machines on the factory floor and their controllers. This traffic is regular (even deterministic for some cases), and has stringent latency requirements in order to arrive within a specified control loop period. By contrast, alarm traffic is sporadic and unpredictable, but nonetheless must be delivered reliably. Each triggered alarm has its strict deadline, until which the alarm can be re-transmitted to reach the BS. The allocation of massive MIMO spatial resources can be optimized for supporting such heterogeneous control and alarm MTC traffic.

Thota *et al.* [97] simulates a mixed traffic of 5% URLLC and 95% non-URLLC devices whose arrivals are modeled by beta and uniform distributions, respectively. To enable the guaranteed access of URLLC devices, a number of preambles can be reserved. Under limited resources, however, this solution may result in a high collision probability of non-URLLC devices. Thus, to increase the preamble resource utilization

TABLE 5. Critical mMTC use cases, requirements, and traffic characteristics [1]–[3]. 3GPP has worked on mMTC/URLLC as an NR approach to meet the stringent reliability and delay needs of critical mMTC use cases, such as manufacturing and robotics.

Use case	Reliability (%)	e2e Latency	Data packet size	Traffic
Smart city: (smart metering, waste management, and smart parking)	99	$\leq$ 5 s	Mostly $\leq$ 200 bytes	Sporadic
Public safety	99.999	1 ms	Small to big	Sporadic
Factory automation (motion control)	99.9999	2 ms	32 bytes	Periodic deterministic
Remote driving	99.999	5 ms	UL: 2.5 Mpbs; Packet size 5220 bytes; DL: 1Mbps; Packet size 2083 bytes	Periodic
Intelligent transport system	99.999	10 ms	UL&DL: 1.1 Mbps; Packet size 1370 bytes	Periodic

of URLLC by exploiting the traffic characteristics (e.g., sporadic URLLC arrivals) is one key enabler.

B. EXISTING SOLUTIONS

To support diverse 5G service types, several network slicing and edge computing solutions have been proposed. These existing solutions are also relevant to supporting critical mMTC, which we will briefly review in this section.

1) RADIO ACCESS NETWORK SLICING

It is generally accepted that network slicing will address various use cases and networking capabilities of the emerging 5G networks. Traditionally, radio access network (RAN) slicing has focused mostly on high layers (e.g., numerology, core network coordination issues in transport and application layers) while addressing their functionality split, computing resource allocation, and coordination. Recently, RAN slicing has also been focusing on radio resource allocation issues in lower layers (e.g., multiple access schemes in MAC layer, interference issues in PHY layer) [7], [8]. Resource allocation is vital for enhancing resource-multiplexing gain between slices while meeting specific RAN slicing service requirements. Regrettably, resource management in RAN slicing is a nontrivial task due to performance isolation, diversified service requirements and network complexities (including mobility and channel statuses). Nonetheless, for 5G RAN slicing, one can build an intelligent resource allocation strategy. The main advantage of scheduling is to reap the benefits of a collaborative learning framework that comprises of deep learning in conjunction with reinforcement learning. For this function, deep learning can be used for large time-scale allocation of resources, while reinforcement learning can be used for online resource management to resolve small-time network dynamics, including imprecision projections and unpredictable network conditions.

2) LOW LATENCY EDGE AND END NODE COMPUTING

Edge computing has been an established paradigm for distributed computing and control services by shifting network load from centralized server towards the last mile networks. At the heart of the 5G wireless systems and beyond is the edge computing capabilities at the cellular BSs. While existing state-of-the-art networks connect and process data centrally (in the cloud or at the edge) for delay and computation-centric applications, both wireless connectivity and database capacity need to be taken closer to the end nodes, thus raising the availability of storage and processing-enabled small cell BSs at the end nodes (e.g., vehicles with BS functionalities [98]).

In addition, the network infrastructure must provide a distributed decision-making service that understands and evolves to cope with the network dynamics to shorten delay. It optimizes communication networks and operations appropriately. One can provide a fresh look to the concept of end node computing by first discussing the applications that the network edge must provide, with a special emphasis on the ensuing challenges in enabling URLLC edge computing services for critical mMTC applications such as VR, V2X, haptics and so forth [76], [99]. Such edge and end node computing paradigm can be integrated with other features (such as software-defined SDNs and NFV) to cope with service disruptions [83], while improving resiliency and reducing latency.

3) SERVICE AWARE SELF-ORGANIZING NETWORKS

While the aforementioned techniques, including RAN slicing, edge and end node computing, and their integrations with SDN and NFV, improve the scalability and flexibility to support diverse service types, it may need to compromise the quality of service and perceived user experiences, due to the integration of multiple techniques and the complications in network configuration management. It is therefore crucial to adopt complementary self-organizing network management strategies, which could provide the desired optimization, distributed management, intelligence, and automation in such convergent network layer approaches [86].

C. CORE NETWORK ARCHITECTURE SOLUTIONS

In addition to improvements in the RAN, radical changes in the core network have been recommended to address the ultra-low delay vision. The innovations in core network introduce new entities such as MEC, SDN, NFV and backhaul approaches [100], [113], [114]. Such improvements are designed to reduce transfer time and overcome the layer-based sequential flow of data as governed by the
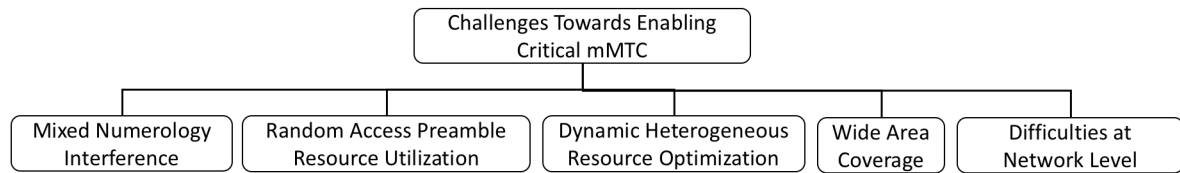
FIGURE 3. Challenges of existing 5G to support critical mMTC communications.

TABLE 6. Summary of core network architectures for URLLC within mMTC.

Architectures	References	Ideas
Edge/Fog-based	[6], [100]–[104]	It minimizes delays, provides near-end computation and storage as well as separates data from its control.
SDN-based	[85], [105], [106]	High performance, massive networking and low delay are envisaged with SDN.
NFV-based	[107]–[112]	NFV decouples the hardware platform dependence and makes it easy to implement suitable features and to share resources within the RAN. It reduces the end-to-end delay by improving the throughput.

existing data transfer protocols. The overall objective is to ensure seamless experience at the utmost possible. Table 6 summarizes the literature currently available on the core network entities which can be used to guarantee low delay for URLLC within mMTC. See [17], [115], [116] and the references therein for details.

VI. CHALLENGES TOWARDS ENABLING CRITICAL mMTC
The main challenges involved in enabling critical mMTC in future cellular networks is shown in Fig. 3 and explained in the following.

A. MIXED-NUMEROLOGY INTERFERENCE
Due to the contrasting requirements of URLLC and mMTC, service configuration differs significantly from the physical layer perspective [13]. In particular, mMTC usually characterized by narrowband transmission with small subcarrier spacing and low baseband sampling rate to support massive connectivity and wide coverage with low-cost and low-power consumption. On the other hand, URLLC usually requires a large subcarrier spacing to meet stringent latency requirement and has high baseband sampling rate. These heterogeneous configuration disparities in baseband and RF inevitably lead to significant interference [117], [118] in critical mMTC.

B. RA PREAMBLE RESOURCE UTILIZATION
In 3GPP, grant-free RA or 2-step RA has been proposed and extensively discussed for both URLLC and mMTC, where the traditional grant request by devices prior to uplink transmission is omitted to reduce the control-signaling overhead and access latency. In grant-free RA, preamble resources are precious and scarce, which needs to be carefully utilized in the context of critical mMTC to balance the requirement trade-off between the URLLC's high access priority and massive access from mMTC. On one hand, preamble reservation (also known as semi-persistent-scheduling (SPS)) for URLLC is preferred to avoid access contention and guarantee its access priority. Nevertheless, in most cases of the coexistence of URLLC and mMTC, it may become resource-utilization inefficient and reduce the preamble availability for accommodating massive sporadic traffic from mMTC. On the other hand, although contention-based grant-free RA is more flexible and efficient in preamble resource utilization, it may cause preamble/data collisions between devices that transmit simultaneously over the shared resource, potentially jeopardizing the transmission reliability and latency. Thus, new challenges are imposed on preamble resource utilization in grant-free RA for critical mMTC, which are to manage preamble collisions across heterogeneous device types while meeting the access requirement of each individual service.

C. DYNAMIC HETEROGENEOUS RESOURCE OPTIMIZATION
In critical mMTC, it is crucial to effectively support data transmissions of both services while taking URLLC as a high priority. Due to limited radio resources, it is essential to accommodate their co-existence by jointly considering their contrasting specifications and requirements in terms of bandwidth, density, latency, and reliability. Thus, how to effectively orchestrate wireless resources in a dynamic and intelligent manner under different levels of service requirements is a challenging task.

D. WIDE AREA COVERAGE (WAC)
Wide areas are common environments in mMTC, but ill-suited for URLLC. This poses challenges to critical mMTC in terms of latency and reliability while making conventional URLLC solutions infeasible. Indeed, preserving massive preambles and increasing transmit power for critical mMTC devices is infeasible subject to limited bandwidth and energy, calling for novel techniques and network architecture design principles.

In addition to time and frequency diversities, spatial diversity needs to be exploited for resolving the constraints in critical mMTC services. However, accurate CSI acquisition
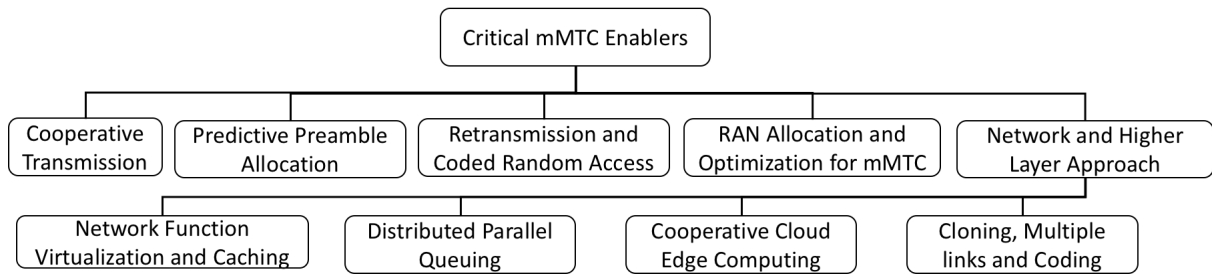
FIGURE 4. Potential enablers for critical mMTC of future cellular IoT networks.

for spatial diversity comes at the cost of feedback overhead which should be reduced to manageable level.

E. DIFFICULTIES AT NETWORK LEVEL

The major difficulties from networking prospective for the critical mMTC can be outlined as follows:

1) NEED FOR HIGHLY FLEXIBLE AND SCALABLE NETWORKING

Owing to the need to flexibly and dynamically serve a great number of dynamically connected devices ranging from a factor of ten to hundred times relative to existing cellular networks, sustaining network efficiency, QoS for URLLC service with increased mMTC network density is a non-trivial challenge.

2) CONGESTION AVOIDANCE AND MANAGEMENT

The incorporation of critical mMTC framework for the existing 5G-based cellular network will result in congestion at different levels of the network including RAN, the core network and the signaling system. As a result, the careful monitoring and management for congestion control in the converged network system is a challenging task.

3) CONTEXT-AWARE DISTRIBUTED COMPUTATION AND CACHING

With the increasing demand of context aware and time sensitive communications in vertical domains, it is essential to investigate synergies among computing and caching resources for critical mMTC as they are distributed across different devices in such massive IoT networks. However, the standard approach to cellular network management is sluggish in terms of network resource management, and a new decentralized approach is required.

VII. POTENTIAL CRITICAL mMTC ENABLERS

Delay-sensitive and delay-tolerant devices can co-exist as discussed in [97], [119], [120]. In this section, we discuss potential enablers for critical mMTC where URLLC (or delay-sensitive) devices and non-URLLC (or delay-tolerant) devices co-exist with different requirements.

A. COOPERATIVE TRANSMISSIONS

Cooperative transmission provides spatial diversity when devices cannot be equipped with multiple antennas. See [121] and the references therein for details. While MTC devices are equipped with single antenna, base stations require high received SNR for critical and reliable communications. Especially, a delay-sensitive device can implement virtual uplink MIMO by achieving the cooperative transmission with certain delay-tolerant devices. Both remaining battery life and geometric distances may be considered in selecting cooperative devices.

B. PREDICTIVE PREAMBLE ALLOCATION

When delay-sensitive and delay-tolerant devices co-exist, preamble allocation to different groups of devices is crucial to meet requirements, especially for delay-sensitive devices. To this end, as in [119], [120], optimal approaches for preamble allocation can be employed for given devices' activities and properties to meet the requirements. However, in practice, it is often difficult to know such information. Thus, new predictive algorithms can be developed.

For example, some delay-sensitive devices may have pseudo-periodic activities that allow a scheduler to predict their activity and perform predictive preamble allocation using learning-based intelligent algorithms. This can result in a low access delay without severe degradation in terms of the spectral efficiency.

C. RE-TRANSMISSION STRATEGIES AND CODED RA

In RA with preamble transmissions, re-transmissions are often inevitable due to a finite number of preambles. To shorten access delay, a short backoff delay can be used for re-transmission strategies. In particular, the backoff delay can be zero and immediate re-transmission can also be allowed [122].

Unfortunately, immediate re-transmission results in the increase of total system load. In other words, the number of URLLC or delay-sensitive devices has to be limited to keep the system stable with guaranteed access delay. To increase the number of devices, the notion of coded RA [123] can be adopted into immediate re-transmission.

D. RAN ALLOCATION AND OPTIMIZATION

A number of research activities have discussed the co-existence of heterogeneous services in a shared physical infrastructure from the RAN resource allocation perspective [8], [124], [125]. Non-orthogonal sharing of RAN resources in uplink communications and heterogeneous NOMA can be developed as potential critical mMTC enablers [8]. By taking advantage of different reliability requirements for different services, reliability diversity can then be introduced as a design principle across the services in order to ensure performance guarantees with non-orthogonal RAN slicing. A context aware risk-sensitive formulation may be considered to allocate resources to the delay sensative devices while minimizing the risk of the mMTC transmission but ensuring URLLC delay/reliability [124]. The URLLC-mMTC co-existence can also be considered by employing power-domain NOMA within a shared resource block, where each sub-carrier can be shared by delay sensitive and delay tolerant devices [125]. Moreover, we may formulate a joint sub-carrier and transmission power allocation problem to maximize the number of successfully connected critical mMTC devices that can satisfy their QoS requirements. Another promising approach with RAN slicing for mMTC and bursty URLLC services may adaptively orchestrate the resources for critical mMTC devices [126]. In addition, a multilevel MAC scheduler can also be investigated to abstract (and intelligently share) the physical resources among network slices [7].

With relevant insights from the aforementioned works and findings discussed earlier, we have outlined the following important ideas for future RAN slicing research.

- *RAN slicing under time-varying channel.* There is a significant trade-off between number of resource blocks and level of modulation and coding scheme - to ensure desired QoS requirements for each (mobile/virtual) network operators. We anticipate that with higher modulation and coding scheme, higher is the spectral efficiency, and may require small number of resource blocks. Such an approach could be useful for plausible handling of the critical mMTC requirements

- *Cross-slice spectrum sharing* will be a promising research direction to improve the RAN slicing-based future Cellular networks. In the standard 5G network, the spectrum management and sharing is usually performed in a homogeneous environment for services of one type only either mMTC or URLLC (and within the slice). However, in the anticipated cross-slice spectrum sharing network, the spectrum sharing mechanism could be more efficient (but highly challenging) when happens across RAN slices facilitating heterogeneous service requirements of critical mMTC applications.

- *Utility optimal theoretic approach* to quantify the price/loss for the network operator while designing RAN slices could offer essential QoS of the critical mMTC even with several network operators and under adverse network conditions, e.g., congested networks with multiple (mobile/virtual) network operators, networks under attacks or failures etc.

- *Adaptive slicing and resource sharing.* As mentioned, existing RAN slice may consist of dedicated and shared resources, e.g., in terms of processing capabilities, memory, and so on and is fully separated from the other network slices and their resources. For critical mMTC applications, in future cellular networks, one may think redesigning an efficient approach to jointly slice and share the resources in an adaptive fashion.

E. NETWORK AND HIGHER LAYERS APPROACH

We propose the following four main ideas for from a higher layer perspectives for the coexistence of delay-sensitive and delay-tolerant devices.

1) NETWORK FUNCTION VIRTUALIZATION AND CACHING FOR RAN

As noted, the service heterogeneity of critical mMTC domains demands diverse QoS requirements. To address the aforementioned challenges of critical mMTC, aggregate network capacity can be abstracted and sliced into multiple virtual networks by exploiting NFV. Moreover, the SDN with NFV not only will facilitate the data and control decoupling but also can provide programmable interface for a network via central controller. With a comprehensive view for the physical network, the controller often enables the efficient utilization of network resources even in time-varying network traffic and channel conditions.

Furthermore, one can divide one physical IoT network into multiple (virtual) networks based on service needs of devices, where the SDN controller can intelligently allocate and update the available system resources among these virtual networks based on their URLLC and/or mMTC requirements. Among all of these virtual networks, each mMTC device can be associated with one virtual network to access the underlying IoT network. In addition to the NFV, it is often beneficial to enhance the aforementioned ideas by using complementary techniques such as intelligent computing and caching from other networks.

2) DISTRIBUTED PARALLEL QUEUING

The standard approaches to enhance the 5G RACH control performance are mainly based on the ALOHA-based approach which are known to suffer from inefficiency, uncertainty and instability issues [28]. One naive approach to this end can be a distributed parallel queuing approach along the lines of that of distributed queuing collision avoidance [127], [128]. It guarantees distributed and stable performing paradigm for critical mMTC. Moreover, multiple distributed queues can operate in parallel, where first queue, can be used for collision resolution for resolving access-request signal collisions, while the other queue, can be used for data transmission and so on for other queues based on the desired set of functionalities. The intrinsic gains due

to distributed parallel queuing will be: i) elimination of the back-off periods and avoidance of collisions in data packet transmissions; ii) independence in performance irrespective of the number of transmitting nodes; iii) stabilization of the traffic conditions; and iv) utilization of few bits for control signaling.

3) COOPERATIVE CLOUD-EDGE COMPUTING

Cloud computing platform has very high temporal and spatial capacity useful for mMTC applications. It provides a comprehensive framework for the underlying network but due to the round trip delay from nodes to the clouds, it may not directly applicable for critical mMTC applications. On the other hand, edge-computing and caching are suitable for URLLC but poses lower computing capacity and memory. To this end, a tightly coupled hybrid cloud and edge computing approach is required for a coordinated and load balanced operation, thus cloud helping edge (and vice versa) satisfying heterogeneous service requirements of critical mMTC. We anticipate that such a cooperative hybrid approach will address the diverse natural challenges of a mix of delay-sensitive and delay-tolerant devices.

4) CLUSTERING WITH DATA SEGREGATION

By categorizing critical mMTC devices into several clusters based on suitable services (such as URLLC and geolocation requirements) and then segregating the device packets at the gateways, the RAN congestion can be ameliorated [129], [130]. In addition, such investigation provides novel insights into energy-efficient clustering along with segregated data sets and may facilitate the implementation of critical NB-IoT and delay-sensitive devices.

5) PACKET CLONING, MULTIPLE CONNECTIVITY AND NETWORK CODING

By considering packet cloning over multiple links as an example, we can always send specially coded packets over several links such that the loss of a coded packet sent over one link can be compensated by another coded packet received from a different link, and the critical mMTC device does not have to wait for a long time to detect and retransmit the lost packet as well. All coded packets will be potentially identical in terms of information content and the numbering and ordering of packets will also become unimportant. Thus, in addition to compensating for the packet losses through data redundancy, controlled cloning and coding can essentially eliminate the notion of sequencing as well. The robustness can be improved not only due to the use of multiple links but also due to the redundancy of the data sent over multiple links.

In fact, network coding has been a celebrated approach for effective handling of decentralized communications and caching services. With respect to critical mMTC applications, network coding can always make it easier for the network operator to improve the efficiency of its data storage against caching, packet transmission or computing nodes failures. Data reliability can be improved substantially with replication, which may be in the form of coded information or basic cloning of packets [131].

Looking forward, conventional block codes, such as the Reed-Solomon codes, can significantly reduce storage costs compared to the storage required by using caching and cloning approach. In general, all information has to be transmitted across the network to retrieve missing information, leading to a memory-repair traffic tradeoff. To this end, network coding can help at optimal points by balancing such memory-repair traffic tradeoff. It will also improve information security and may support memory-repair trading despite other expensive approach at the cellular infrastructure.

VIII. CONCLUDING REMARKS

It is anticipated that next generation cellular IoT networks will support massive connectivity for resource-constrained devices while meeting their diverse and critical QoS requirements. We need to address a number of challenges in order to improve performance such as scalability, reliability and latency. Within this article, we include a literature review of mMTC and URLLC to enable essential mMTC. To this end, critical mMTC is characterized by mMTC to support both delay-sensitive (or URLLC devices) and delay-tolerant devices. Since two different types of devices co-exist in a system, we have studied diverse requirements for heterogeneous features of critical mMTC in this paper. In particular, existing approaches have been reviewed and key enabling technologies from different perspectives are also identified.

Along with the growing presence of 5G in various corners of the world, a few research groups have started questioning: What will 6G going to be? The 6G vision requires extensive future investigations, including data-driven service, which could potentially enable (human-level) instant and intelligent wireless connectivity. The significant foundation for 6G innovation is anticipated with seamless critical mMTC and URLLC services, powered by the impetus to provide vertical-specific solutions.

REFERENCES

[1] 5G Americas. (2020). *New Services and Applications With 5G Ultra-Reliable Low-Latency Communication*. [Online]. Available: https://www.5gamericas.org/new-services-applications-with-5g-ultra-reliable-low-latency-communications/

[2] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC) (Release 16)*, document 3GPP TR 38.824 V2.0.1, 2019.

[3] Y. Li, X. Cheng, Y. Cao, D. Wang, and L. Yang, "Smart choice for the smart grid: Narrowband Internet of Things (NB-IoT)," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1505–1515, Jun. 2018.

[4] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," 2019, *arXiv:1902.10265*. [Online]. Available: http://arxiv.org/abs/1902.10265

[5] R. Trivisonno, R. Guerzoni, I. Vaishnavi, and D. Soldani, "Towards zero latency software defined 5G networks," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2566–2571.

[6] D. Feng, C. She, K. Ying, L. Lai, Z. Hou, T. Q. S. Quek, Y. Li, and B. Vucetic, "Toward ultrareliable low-latency communications: Typical scenarios, possible solutions, and open issues," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 94–102, Jun. 2019.

[7] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[8] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[9] J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, and M. Bennis, "Extreme URLLC: Vision, challenges, and key enablers," 2020, *arXiv:2001.09683*. [Online]. Available: http://arxiv.org/abs/2001.09683

[10] N. H. Mahmood, S. Böcker, A. Munari, F. Clazzer, I. Moerman, K. Mikhaylov, O. Lopez, O. S. Park, E. Mercier, H. Bartz, and R. Jüntti, "White paper on critical and massive machine type communication towards 6G," 2020, *arXiv:2004.14146*. [Online]. Available: http://arxiv.org/abs/2004.14146

[11] S. K. Sharma, I. Woungang, A. Anpalagan, and S. Chatzinotas, "Toward tactile Internet in beyond 5G era: Recent advances, current issues, and future directions," *IEEE Access*, vol. 8, pp. 56948–56991, 2020.

[12] 5G Americas. (2020). *The 5G Evolution: 3GPP Releases 16-17*. [Online]. Available: https://www.5gamericas.org/5g-evolution-3gpp-Releases-16-17/

[13] A. B. Kihero, M. Sohaib J. Solaija, and H. Arslan, "Multi-numerology multiplexing and inter-numerology interference analysis for 5G," 2019, *arXiv:1905.12748*. [Online]. Available: http://arxiv.org/abs/1905.12748

[14] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 2015.

[15] H. Wang and A. O. Fapojuwo, "A survey of enabling technologies of low power and long range Machine-to-Machine communications," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2621–2639, 4th Quart., 2017.

[16] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[17] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3098–3130, 4th Quart., 2018.

[18] A. Hoglund, X. Lin, O. Liberg, A. Behravan, E. A. Yavuz, M. Van Der Zee, Y. Sui, T. Tirronen, A. Ratilainen, and D. Eriksson, "Overview of 3GPP release 14 enhanced NB-IoT," *IEEE Netw.*, vol. 31, no. 6, pp. 16–22, Nov. 2017.

[19] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[20] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-reliable low latency cellular networks: Use cases, challenges and approaches," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 119–125, Dec. 2018.

[21] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE–A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, 2nd Quart., 2015.

[22] M. Elsaadany, A. Ali, and W. Hamouda, "Cellular LTE–A technologies for the future Internet-of-Things: Physical layer features and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2544–2572, 4th Quart., 2017.

[23] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[24] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5g: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.

[25] A. Rico-Alvarino, M. Vajapeyam, H. Xu, X. Wang, Y. Blankenship, J. Bergman, T. Tirronen, and E. Yavuz, "An overview of 3GPP enhancements on machine to machine communications," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 14–21, Jun. 2016.

[26] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2020.

[27] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[28] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE–A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.

[29] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE–A random access for massive machine-type communications in smart cities," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 76–83, Jan. 2017.

[30] W. Yang, M. Wang, J. Zhang, J. Zou, M. Hua, T. Xia, and X. You, "Narrowband wireless access for low-power massive Internet of Things: A bandwidth perspective," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 138–145, Jun. 2017.

[31] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free random access solutions for URLLC services in 5G new radio," 2019, *arXiv:1904.05593*. [Online]. Available: http://arxiv.org/abs/1904.05593

[32] H. Elhammouti, E. Sabir, M. Benjillali, L. Echabbi, and H. Tembine, "Self-organized connected objects: Rethinking QoS provisioning for IoT services," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 41–47, 2017.

[33] S. Verma, Y. Kawamoto, Z. M. Fadlullah, H. Nishiyama, and N. Kato, "A survey on network methodologies for real-time analytics of massive IoT data and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1457–1477, 3rd Quart., 2017.

[34] R. Atat, L. Liu, J. Wu, G. Li, C. Ye, and Y. Yang, "Big data meet cyber-physical systems: A panoramic survey," *IEEE Access*, vol. 6, pp. 73603–73636, 2018.

[35] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[36] X. Wang, X. Li, and V. C. M. Leung, "Artificial intelligence-based techniques for emerging heterogeneous network: State of the arts, opportunities, and challenges," *IEEE Access*, vol. 3, pp. 1379–1391, 2015.

[37] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in Internet of Things: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 1–27, Feb. 2018.

[38] G. Yu, R. Liu, Q. Chen, and Z. Tang, "A hierarchical SDN architecture for ultra-dense millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 79–85, Jun. 2018.

[39] S. Andreev, V. Petrov, M. Dohler, and H. Yanikomeroglu, "Future of ultra-dense networks beyond 5G: Harnessing heterogeneous moving cells," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 86–92, Jun. 2019.

[40] *Study on Scenarios and Requirements for Next Generation Access Technologies (Rel-14)*, document TR 38.913, v14.3.0, 2017.

[41] R. Ratasuk, N. Mangalvedhe, Y. Zhang, M. Robert, and J.-P. Koskinen, "Overview of narrowband IoT in LTE rel-13," in *Proc. IEEE Conf. Standards for Commun. Netw. (CSCN)*, Oct. 2016, pp. 1–7.

[42] *Evolved Universal Terrestrial Radio Access; Physical Layer Procedures (Rel-14)*, document TS 36.213, 3GPP, V14.5.0, 2018.

[43] *Evolved universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, document TS 36.211, 3GPP, V15.5.0, 2019. [Online]. Available: ftp://ftp.3gpp.org/specs/2019-03/Rel-15/36 series/

[44] X. Lin, A. Adhikary, and Y.-P. E. Wang, "Random access preamble design and detection for 3GPP narrowband IoT systems," *IEEE Wireless Commun. Lett.*, vol. 5, no. 6, pp. 640–643, Dec. 2016.

[45] Y. D. Beyene, R. Jantti, K. Ruttik, and S. Iraji, "On the performance of narrow-band Internet of Things (NB-IoT)," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[46] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz, and J. M. Lopez-Soler, "An analytical performance evaluation framework for NB-IoT," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7232–7240, Aug. 2019.

[47] R. Ratasuk, N. Mangalvedhe, Z. Xiong, M. Robert, and D. Bhatoolaul, "Enhancements of narrowband IoT in 3GPP rel-14 and rel-15," in *Proc. IEEE Conf. Standards for Commun. Netw. (CSCN)*, Sep. 2017, pp. 60–65.

[48] J. Lee and J. Lee, "Prediction-based energy saving mechanism in 3GPP NB-IoT networks," *Sensors*, vol. 17, no. 9, p. 2008, Sep. 2017.

[49] J. Zhang, D. Xie, X. Wang, and J. Lu, "Exploiting propagation delay difference in collided preambles for efficient random access in NB-IoT," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[50] T. Kim, D. M. Kim, N. Pratas, P. Popovski, and D. K. Sung, "An enhanced access reservation protocol with a partial preamble transmission mechanism in NB-IoT systems," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2270–2273, Oct. 2017.

[51] F. Wu, B. Zhang, W. Fan, X. Tian, S. Huang, C. Yu, and Y. Liu, "An enhanced random access algorithm based on the clustering-reuse preamble allocation in NB-IoT system," *IEEE Access*, vol. 7, pp. 183847–183859, 2019.

[52] W. S. Jeon, S. B. Seo, and D. G. Jeong, "Effective frequency hopping pattern for ToA estimation in NB-IoT random access," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 10150–10154, Oct. 2018.

[53] R. Harwahyu, R.-G. Cheng, C.-H. Wei, and R. F. Sari, "Optimization of random access channel in NB-IoT," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 391–402, Feb. 2018.

[54] S. Ha, H. Seo, Y. Moon, D. Lee, and J. Jeong, "A novel solution for NB-IoT cell coverage expansion," in *Proc. Global Internet Things Summit (GIoTS)*, Jun. 2018, pp. 1–5.

[55] S. Ravi, P. Zand, M. El Soussi, and M. Nabi, "Evaluation, modeling and optimization of coverage enhancement methods of NB-IoT," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–7.

[56] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Analytic analysis of narrowband IoT coverage enhancement approaches," in *Proc. Global Internet Things Summit (GIoTS)*, Jun. 2018, pp. 1–6.

[57] P.-Y. Liu, K.-R. Wu, J.-M. Liang, J.-J. Chen, and Y.-C. Tseng, "Energy-efficient uplink scheduling for ultra-reliable communications in NB-IoT networks," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–5.

[58] A. Azari, C. Stefanović, P. Popovski, and C. Cavdar, "On the latency-energy performance of NB-IoT systems in providing wide-area IoT connectivity," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 57–68, Mar. 2020.

[59] C.-Y. Chen, A. C.-S. Huang, S.-Y. Huang, and J.-Y. Chen, "Energy-saving scheduling in the 3GPP narrowband Internet of Things (NB-IoT) using energy-aware machine-to-machine relays," in *Proc. 27th Wireless Opt. Commun. Conf. (WOCC)*, Apr. 2018, pp. 1–3.

[60] J. Oh and H. Song, "Study on the effect of LTE on the coexistence of NB-IoT," in *Proc. 10th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2018, pp. 610–612.

[61] G. J. González, F. H. Gregorio, and J. Cousseau, "Interference analysis in the LTE and NB-IoT uplink multiple access with RF impairments," in *Proc. IEEE 23rd Int. Conf. Digit. Signal Process. (DSP)*, Nov. 2018, pp. 1–4.

[62] S. Liu, L. Xiao, Z. Han, and Y. Tang, "Eliminating NB-IoT interference to LTE system: A sparse machine learning-based approach," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6919–6932, Aug. 2019.

[63] B. Yang, L. Zhang, D. Qiao, G. Zhao, and M. A. Imran, "Narrowband Internet of Things (NB-IoT) and LTE systems co-existence analysis," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[64] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: A view on 5G cellular technology beyond 3GPP release 15," *IEEE Access*, vol. 7, pp. 127639–127651, 2019.

[65] P. Yang, X. Xi, T. Q. S. Quek, J. Chen, X. Cao, and D. Wu, "RAN slicing for massive IoT and bursty URLLC service multiplexing: Analysis and optimization," 2020, *arXiv:2001.04161*. [Online]. Available: http://arxiv.org/abs/2001.04161

[66] T. Taleb, I. Afolabi, and M. Bagaa, "Orchestrating 5G network slices to support industrial Internet and to shape next-generation smart factories," *IEEE Netw.*, vol. 33, no. 4, pp. 146–154, Jul. 2019.

[67] S. Din, A. Paul, and A. Rehman, "5G-enabled hierarchical architecture for software-defined intelligent transportation system," *Comput. Netw.*, vol. 150, pp. 81–89, Feb. 2019.

[68] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 106984.

[69] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 607–612.

[70] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Joint resource configuration and MCS selection scheme for uplink grant-free URLLC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[71] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," 2020, *arXiv:2002.07842*. [Online]. Available: https://arxiv.org/abs/2002.07842

[72] J. Choi, "Repetition-based NOMA transmission and its outage probability analysis," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 5913–5922, Jun. 2020.

[73] M. Maule, P.-V. Mekikis, K. Ramantas, J. Vardakas, and C. Verikoukis, "Real-time dynamic network slicing for the 5G radio access network," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[74] E. J. dos Santos, R. D. Souza, J. L. Rebelatto, and H. Alves, "Network slicing for URLLC and eMBB with max-matching diversity channel allocation," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 658–661, Mar. 2020.

[75] F. Song, J. Li, C. Ma, Y. Zhang, L. Shi, and D. N. K. J. Li, "Dynamic virtual resource allocation for 5G and beyond network slicing," 2020, *arXiv:2004.00506*. [Online]. Available: http://arxiv.org/abs/2004.00506

[76] A. Ksentini and P. A. Frangoudis, "Toward slicing-enabled multi-access edge computing in 5G," *IEEE Netw.*, vol. 34, no. 2, pp. 99–105, Mar. 2020.

[77] S. O. Oladejo and O. E. Falowo, "Latency-aware dynamic resource allocation scheme for multi-tier 5G network: A network slicing-multitenancy scenario," *IEEE Access*, vol. 8, pp. 74834–74852, 2020.

[78] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6419–6432, Oct. 2018.

[79] E. Peltonen, M. Bennis, M. Capobianco, M. Debbah, A. Ding, F. Gil-Castiñeira, M. Jurmu, T. Karvonen, M. Kelanti, A. Kliks, T. Leppänen, L. Lovén, T. Mikkonen, A. Rao, S. Samarakoon, K. Seppänen, P. Sroka, S. Tarkoma, and T. Yang, "6G white paper on edge intelligence," 2020, *arXiv:2004.14850*. [Online]. Available: http://arxiv.org/abs/2004.14850

[80] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, "Network slicing games: Enabling customization in multi-tenant mobile networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 662–675, Apr. 2019.

[81] A. Ksentini, P. A. Frangoudis, A. Pc, and N. Nikaein, "Providing low latency guarantees for slicing-ready 5G systems via two-level MAC scheduling," *IEEE Netw.*, vol. 32, no. 6, pp. 116–123, Nov. 2018.

[82] J. Zheng, G. de Veciana, and A. Banchs, "Constrained network slicing games: Achieving service guarantees and network efficiency," 2020, *arXiv:2001.01402*. [Online]. Available: http://arxiv.org/abs/2001.01402

[83] R. Ford, A. Sridharan, R. Margolies, R. Jana, and S. Rangan, "Provisioning low latency, resilient mobile edge clouds for 5G," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 169–174.

[84] C.-F. Liu, M. Bennis, and H. V. Poor, "Latency and reliability-aware task offloading and resource allocation for mobile edge computing," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–7.

[85] K. Sood, S. R. Pokhrel, K. Karmakar, V. Vardharajan, and S. Yu, "SDN-capable IoT last-miles: Design challenges," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[86] 5G PPP Architecture Working Group, "View on 5G architecture (Version 3.0)," 5G PPP Archit. Work. Group, White Paper, Jun. 2019.

[87] K. Zhang, X. Xu, J. Zhang, B. Zhang, X. Tao, and Y. Zhang, "Dynamic multiconnectivity based joint scheduling of eMBB and uRLLC in 5G networks," *IEEE Syst. J.*, early access, Apr. 6, 2020, doi: 10.1109/JSYST.2020.2977666.

[88] J. Yang, "5G network for a variety of vertical services," in *Proc. 5G Verticals*, 2020, pp. 25–56.

[89] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," *IEEE Netw.*, vol. 34, no. 2, pp. 186–194, Mar. 2020.

[90] D. Han and W. Chen, "QoS differential scheduling of URLLC under FIFO service discipline: A cross-layer approach," *IEEE Wireless Commun. Lett.*, early access, Apr. 28, 2020, doi: 10.1109/LWC.2020.2990676.

[91] D. Han, W. Chen, and Y. Fang, "Joint channel and queue aware scheduling for latency sensitive mobile edge computing with power constraints," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3938–3951, Jun. 2020.

[92] B. Collins and R. L. Cruz, "Transmission policies for time varying channels with average delay constraints," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, 1998, pp. 709–717.

[93] C. Wang, B. Li, K. Sohraby, M. Daneshmand, and Y. Hu, "Upstream congestion control in wireless sensor networks through cross-layer optimization," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 786–795, May 2007.

[94] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[95] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Preamble reservation based access for grouped mMTC devices with URLLC requirements," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[96] E. Fitzgerald and M. Pióro, "Efficient pilot allocation for URLLC traffic in 5G industrial IoT networks," in *Proc. 11th Int. Workshop Resilient Netw. Design Model. (RNDM)*, Oct. 2019, pp. 1–7.

[97] J. Thota and A. Aijaz, "On performance evaluation of random access enhancements for 5G uRLLC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–7.

[98] S. R. Pokhrel and J. Choi, "Improving TCP performance over WiFi for Internet of Vehicles: A federated learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6798–6802, Jun. 2020.

[99] M. S. Elbamby, C. Perfecto, C.-F. Liu, J. Park, S. Samarakoon, X. Chen, and M. Bennis, "Wireless edge computing with latency and reliability guarantees," *Proc. IEEE*, vol. 107, no. 8, pp. 1717–1737, Aug. 2019.

[100] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[101] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," *ETSI White Paper*, vol. 11, no. 11, pp. 1–16, 2015.

[102] Y. Zhang, B. Di, P. Wang, J. Lin, and L. Song, "HetMEC: Heterogeneous multi-layer mobile edge computing in the 6G era," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4388–4400, Apr. 2020.

[103] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 22–29, Dec. 2016.

[104] L. Nkenyereye, C. H. Liu, and J. Song, "Towards secure and privacy preserving collision avoidance system in 5G fog based Internet of Vehicles," *Future Gener. Comput. Syst.*, vol. 95, pp. 488–499, Jun. 2019.

[105] D. A. Chekired, M. A. Togou, L. Khoukhi, and A. Ksentini, "5G-slicing-enabled scalable SDN core network: Toward an ultra-low latency of autonomous driving service," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1769–1782, Aug. 2019.

[106] S. Panev and P. Latkoski, "SDN-based failure detection and recovery mechanism for 5G core networks," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 2, p. e3721, Feb. 2020.

[107] X. Ge, R. Zhou, and Q. Li, "5G NFV-based tactile Internet for mission-critical IoT services," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6150–6163, Jul. 2020.

[108] G. C. Valastro, D. Panno, and S. Riolo, "A SDN/NFV based C-RAN architecture for 5G mobile networks," in *Proc. Int. Conf. Sel. Topics Mobile Wireless Netw. (MoWNeT)*, Jun. 2018, pp. 1–8.

[109] H. Chantre and N. L. S. D. Fonseca, "Reliable broadcasting in 5G NFV-based networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 218–224, Mar. 2018.

[110] H. D. Chantre and N. L. S. da Fonseca, "Multi-objective optimization for edge device placement and reliable broadcasting in 5G NFV-based small cell networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2304–2317, Oct. 2018.

[111] Z. Shu and T. Taleb, "A novel QoS framework for network slicing in 5G and beyond networks based on SDN and NFV," *IEEE Netw.*, vol. 34, no. 3, pp. 256–263, May 2020.

[112] H. A. Alameddine, C. Assi, M. H. Kamal Tushar, and J. Y. Yu, "Low-latency service schedule orchestration in NFV-based networks," in *Proc. IEEE Conf. Netw. Softw. (NetSoft)*, Jun. 2019, pp. 378–386.

[113] V.-G. Nguyen, A. Brunstrom, K.-J. Grinnemo, and J. Taheri, "SDN/NFV-based mobile packet core network architectures: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1567–1602, 3rd Quart., 2017.

[114] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[115] Ericsson, "5G systems—Enabling the transformation of industry and society," Ericsson, Stockholm, Sweden, White Paper, Jan. 2017. [Online]. Available: https://www.ericsson.com/49daeb/assets/local/reports-papers/white-papers/wp-5g-systems.pdf

[116] S. Ludwig et al., "A 5G architecture for the factory of the future," in *Proc. IEEE 23rd Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, vol. 1, Sep. 2018, pp. 1409–1416.

[117] A. Yazar and H. Arslan, "Flexible multi-numerology systems for 5G new radio," *J. Mobile Multimedia*, vol. 14, no. 4, pp. 367–394, 2018.

[118] L. Zhang, A. Ijaz, P. Xiao, and R. Tafazolli, "Channel equalization and interference analysis for uplink narrowband Internet of Things (NB-IoT)," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2206–2209, Oct. 2017.

[119] N. Li, C. Cao, and C. Wang, "Dynamic resource allocation and access class barring scheme for delay-sensitive devices in machine to machine (M2M) communications," *Sensors*, vol. 17, no. 6, p. 1407, Jun. 2017.

[120] C. Zhang, X. Sun, J. Zhang, X. Wang, S. Jin, and H. Zhu, "Throughput optimization with delay guarantee for massive random access of M2M communications in industrial IoT," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10077–10092, Dec. 2019.

[121] Y. Cao, B. Vojcic, and M. Souryal, "User-cooperative transmission with channel feedback in slow fading environment," in *Proc. IEEE 60th Veh. Technol. Conf. (VTC-Fall)*, vol. 3, Sep. 2004, pp. 2063–2067.

[122] J. Choi, "Low-latency multichannel ALOHA with fast retrial for machine-type communications," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3175–3185, Apr. 2019.

[123] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.

[124] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, "EMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.

[125] A. E. Mostafa, Y. Zhou, and V. W. S. Wong, "Connectivity maximization for narrowband IoT systems with NOMA," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[126] B. Yang, L. Zhang, O. Onireti, P. Xiao, M. A. Imran, and R. Tafazolli, "Mixed-numerology signals transmission and interference cancellation for radio access network slicing," *IEEE Trans. Wireless Commun.*, early access, Apr. 28, 2020, doi: 10.1109/TWC.2020.2989399.

[127] E. Kartsakli, A. Cateura, L. Alonso, J. Alonso-Zarate, and C. Verikoukis, "Cross-layer enhancement for wlan systems with heterogeneous traffic based on DQCA," *IEEE Commun. Mag.*, vol. 46, no. 6, pp. 60–66, Jun. 2008.

[128] L. Alonso, R. Ferrus, and R. Agusti, "WLAN throughput improvement via distributed queuing MAC," *IEEE Commun. Lett.*, vol. 9, no. 4, pp. 310–312, Apr. 2005.

[129] Y. Sim and D.-H. Cho, "Performance evaluation of partially clustered access scheme for massive machine type communications," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 626–629, Mar. 2020.

[130] Z. Li, J. Chen, R. Ni, S. Chen, X. Li, and Q. Zhao, "Enabling heterogeneous mMTC by energy-efficient and connectivity-aware clustering and routing," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[131] S. R. Pokhrel and J. Choi, "Low-delay scheduling for Internet of Vehicles: Load-balanced multipath communication with FEC," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8489–8501, Dec. 2019.
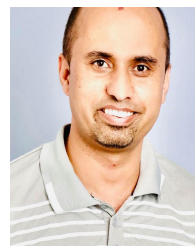
SHIVA RAJ POKHREL (Member, IEEE) received the B.E. and M.E. degrees from Pokhara University, Nepal, in 2007 and 2013, respectively, and the Ph.D. degree from the Swinburne University of Technology, Australia, in 2017. He was a Research Fellow with The University of Melbourne and a Network Engineer with Nepal Telecom, from 2007 to 2014. He is currently a Lecturer of mobile computing with Deakin University, Geelong, VIC, Australia. His research interests include multiconnectivity, federated learning, industry 4.0 automation, blockchain modeling, optimization, recommender systems, 6G, cloud computing, dynamics control, the Internet of Things, cyber-physical systems and their applications in smart manufacturing, and autonomous vehicles and cities. He served as the Workshop Chair/Publicity Co-Chair for the several IEEE/ACM conferences, including the IEEE INFOCOM, the IEEE GLOBECOM, the IEEE ICC, ACM MobiCom, and so on. He was a recipient of the Prestigious Marie Skłodowska-Curie Grant Fellowship, in 2017.

JIE DING received the Ph.D. degree in communication engineering from Macquarie University, Sydney, NSW, Australia, in 2016. He is currently with the School of Information Technology, Deakin University, Burwood, VIC, Australia. His research interests include machine-type communication, massive MIMO, and random access.

JIHONG PARK (Member, IEEE) received the B.S. and Ph.D. degrees from Yonsei University, Seoul, South Korea, in 2009 and 2016, respectively. He was a Visiting Researcher with The Hong Kong Polytechnic University, KTH, Sweden, Aalborg University, and the New Jersey Institute of Technology, USA, in 2013, 2015, 2016, and 2017, respectively. From 2016 to 2017, he was a Postdoctoral Researcher with Aalborg University, Denmark. He was a Postdoctoral Researcher with the University of Oulu, Finland. Since 2020, he has been a Lecturer with the School of Information Technology, Deakin University, Australia. His research interests include distributed machine learning and ultrareliable designs in 5G communication systems and beyond. He received the 2014 IEEE GLOBECOM Student Travel Grant, the 2014 IEEE Seoul Section Student Paper Contest Bronze Prize, and the Sixth IDIS-ETNEWS (The Electronic Times) Paper Contest Award sponsored by the Ministry of Science, ICT, and Future Planning, South Korea.

OK-SUN PARK (Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Chonnam National University, South Korea, in 1996 and 1999, respectively, and the Ph.D. degree in information and communication engineering from Chungnam National University, South Korea, in 2014. She has been with the Electronics and Telecommunications Research Institute, as a Principal Researcher, since 1999. Her research interests include channel coding and medium access technologies for B5G/6G communication systems.
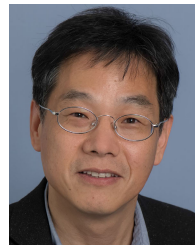
JINHO CHOI (Senior Member, IEEE) was born in Seoul, South Korea. He received the B.E. degree *(magna cum laude)* in electronics engineering from Sogang University, Seoul, in 1989, and the M.S.E. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1991 and 1994, respectively. He was with Swansea University, U.K., as a Professor/Chair in wireless. He was also with the Gwangju Institute of Science and Technology (GIST), South Korea, as a Professor. He is currently with the School of Information Technology, Deakin University, Burwood, VIC, Australia, as a Professor. He has authored two books published by the Cambridge University Press, in 2006 and 2010. His research interests include the Internet of Things (IoT), wireless communications, and statistical signal processing. He received the 1999 Best Paper Award for Signal Processing from EURASIP and the 2009 Best Paper Award from WPMC Conference. He serves as an Editor for the IEEE Transactions on Communications and the IEEE Wireless Communications Letters. He also serves as a Division Editor for the *Journal of Communications and Networks (JCN)*. He served as an Associate Editor or Editor for other journals, including the IEEE Communications Letters, *JCN*, the IEEE Transactions on Vehicular Technology, and *ETRI Journal*.

● ● ●